(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
7. Juni 2001 (07.06.2001)

PCT

(10) Internationale Veröffentlichungsnummer
WO 01/40510 A2

(51) Internationale Patentklassifikation[7]: C12Q 1/68

(21) Internationales Aktenzeichen: PCT/EP00/11978

(22) Internationales Anmeldedatum:
29. November 2000 (29.11.2000)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
199 57 320.4 29. November 1999 (29.11.1999) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: **FEBIT FERRARIUS BIOTECHNOLOGY GMBH** [DE/DE]; Käfertalerstrasse 190, 68167 Mannheim (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder *(nur für US)*: **KAUSCH, Andrea** [DE/DE]; Ricarda-Huch-Strasse 3, 64291 Darmstadt (DE). **STÄHLER, Cord, F.** [DE/DE]; Siegfriedstrasse 9, 69469 Weinheim (DE). **STÄHLER, Peer, F.** [DE/DE]; Riedfeldstrasse 3, 68169 Mannheim (DE). **BAUM, Michael** [DE/DE]; Albert-Fritz-Strasse 74, 69124 Heidelberg (DE). **MÜLLER, Manfred** [DE/DE]; Reutterstrasse 76b, 80689 München (DE).

(74) Anwälte: **WEICKMANN & WEICKMANN** usw.; Postfach 860 820, 81635 München (DE).

(81) Bestimmungsstaaten *(national)*: AU, CA, JP, US.

(84) Bestimmungsstaaten *(regional)*: europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Veröffentlicht:
— *Ohne internationalen Recherchenbericht und erneut zu veröffentlichen nach Erhalt des Berichts.*

Zur Erklärung der Zweibuchstaben-Codes, und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

(54) Title: DYNAMIC SEQUENCING BY HYBRIDIZATION

(54) Bezeichnung: DYNAMISCHE SEQUENZIERUNG DURCH HYBRIDISIERUNG

(57) Abstract: The invention relates to a method for sequencing nucleic acids using carrier chips that contain polymer probes, which are constructed of nucleotides and/or nucleotide analogs and which permit a specific binding with nucleic acids present in the sample. The method is dynamically carried out in a number of cycles, whereby the sequence information obtained from a preceding cycle is used for modifying carrier-bound probes in the subsequent cycle.

(57) Zusammenfassung: Die Erfindung betrifft ein Verfahren zur Sequenzierung von Nukleinsäuren unter Verwendung von Trägerchips, die aus Nukleotiden oder/und Nukleotideanaloga aufgebaute Polymersonden enthalten und eine spezifische Bindung mit in einer Probe vorhandenen Nukleinsäuren erlauben. Das Verfahren wird dynamisch in mehreren Zyklen durchgeführt, wobei die aus einem vorhergehenden Zyklus gewonnenen Sequenzinformationen zur Modfizierung trägergebundener Sonden im nachfolgenden Zyklus genutzt werden.

WO 01/40510 A2

Dynamische Sequenzierung durch Hybridisierung

Beschreibung

Die Erfindung betrifft ein Verfahren zur Sequenzierung von Nukleinsäuren unter Verwendung von Trägerchips, die aus Nukleotiden oder/und Nukleotideanaloga aufgebaute Polymersonden enthalten und eine spezifische Bindung mit in einer Probe vorhandenen Nukleinsäuren erlauben. Das Verfahren wird dynamisch in mehreren Zyklen durchgeführt, wobei die aus einem vorhergehenden Zyklus gewonnenen Sequenzinformationen zur Modfizierung trägergebundener Sonden im nachfolgenden Zyklus genutzt werden.

1. Einleitung

Für die Grundlagenforschung, die Medizin, die Biotechnologie sowie weitere wissenschaftliche Disziplinen ist die Erfassung biologisch relevanter Information in definiertem Untersuchungsmaterial von herausragender Bedeutung. Zumeist steht dabei die genetische Information im Mittelpunkt des Interesses. Diese genetische Information besteht in einer enormen Vielfalt unterschiedlicher Nukleinsäuresequenzen, der DNA. Die Nutzung dieser Information im biologischen Organismus führt über die Herstellung von Abschriften der DNA in RNA meist zur Synthese von Proteinen.

Um diese Wirkprinzipien der Natur besser verstehen zu können, ist eine effiziente und sichere Entschlüsselung von DNA-Sequenzen notwendig. Die Detektion von Nukleinsäuren und die Bestimmung der Abfolge der vier Basen in der Kette der Nukleotide, die generell als Sequenzierung bezeichnet wird, liefert wertvolle Daten für Forschung und angewandte Medizin. In der Medizin konnte in stark zunehmendem Maße durch die in vitro-Diagnostik (IVD) ein Instrumentarium zur Bestimmung wichtiger Patientenparameter

entwickelt und dem behandelnden Arzt zur Verfügung gestellt werden. Für viele Erkrankungen wäre eine Diagnose zu einem ausreichend frühen Zeitpunkt ohne dieses Instrumentarium nicht möglich. Hier hat sich die genetische Analyse als wichtiges neues Verfahren etabliert.

In enger Verzahnung von Grundlagenforschung und klinischer Forschung konnten die molekularen Ursachen und (pathologischen) Zusammenhänge einiger Krankheitsbilder bis auf die Ebene der genetischen Information zurückverfolgt und aufgeklärt werden. Diese wissenschaftliche Vorgehensweise steht jedoch noch am Anfang ihrer Entwicklung und gerade für ihre Umsetzung im Rahmen von Therapiestrategien bedarf es stark intensivierter Anstrengungen. Insgesamt haben die Genomwissenschaften und die damit im Zusammenhang stehende Nukleinsäureanalytik sowohl zum Verständnis der molekularen Grundlagen des Lebens als auch zur Aufklärung sehr komplexer Krankheitsbilder und pathologischer Vorgänge wichtige Beiträge geleistet.

2. Stand der Technik

Genetische Information wird durch Analyse von Nukleinsäuren, meist in Form von DNA, gewonnen. Es gibt drei wesentliche Techniken für die Analyse von DNA. Die erste wird als Polymerase-Kettenreaktion (PCR) bezeichnet. Diese und verwandte Methoden dienen der selektiven enzymgestützen Vervielfältigung (Amplifikation) von DNA, indem kurze flankierende DNA Stränge mit bekannter Sequenz genutzt werden, um die enzymatische Synthese des dazwischen liegenden Bereiches zu starten. Dabei muß die Sequenz dieses Bereiches nicht im Detail bekannt sein. Der Mechanismus erlaubt damit anhand eines kleinen Ausschnittes an Information (den flankierenden DNA Strängen) die selektive Vervielfältigung eines bestimmten DNA Abschnittes, so daß dieser vervielfältigte DNA Strang in großer Menge für weitere Arbeiten und Analysen zur Verfügung steht.

Als zweite Basistechnik wird die Elektrophorese verwendet. Dabei handelt es sich um eine Technik zur Trennung von DNA Molekülen anhand ihrer Größe. Die Trennung erfolgt in einem elektrischen Feld, das die DNA Moleküle zur Wanderung zwingt. Durch geeignete Medien, wie z.B. vernetzte Gele, wird die Bewegung im elektrischen Feld abhängig von der Molekülgröße erschwert, so daß kleine Moleküle und damit kürzere DNA Fragmente schneller wandern als längere. Elektrophorese ist die wichtigste etablierte Methode für die DNA Sequenzierung und darüber hinaus für viele Verfahren zur Reinigung und Analyse von DNA. Das verbreitetste Verfahren ist die Flachbett-Gelelektrophorese, die im Bereich der Hochdurchsatzsequenzierung allerdings zunehmend von der Kapillar-Gelelektrophorese verdrängt wird.

Bei der dritten Methode handelt es sich um die Analyse von Nukleinsäuren durch sogenannte Hybridisierung. Hierbei wird eine DNA-Sonde mit bekannter Sequenz verwendet, um eine komplementäre Nukleinsäure zu identifizieren, meistens vor dem Hintergrund eines komplexen Gemisches von sehr vielen DNA- oder RNA-Molekülen. Die passenden Stränge binden sich stabil und sehr spezifisch aneinander.

Die drei Basistechniken kommen häufig in Kombination vor, indem z.B. das Probenmaterial für ein Hybridisierungsexperiment vorher selektiv durch PCR vervielfältigt wird.

Bei der Sequenzanalyse auf einem DNA-Trägerchip nutzt man ebenfalls das Prinzip der Hybridisierung von zueinander passenden DNA-Strängen aus. Die Entwicklung von DNA-Trägerchips oder DNA-Arrays bedeutet eine extreme Parallelisierung und Miniaturisierung des Formats von Hybridisierungs-experimenten. DNA in einer Probe kann nur an den Stellen an die auf dem Träger fixierte DNA binden, an denen die Sequenz der beiden DNA-Stränge übereinstimmt. Mit Hilfe der fixierten DNA auf dem Träger kann selektiv die komplementäre DNA in der Probe nachgewiesen werden. Dadurch werden

beispielsweise Mutationen im Probenmaterial durch das Muster erkannt, das nach der Hybridisierung auf dem Träger entsteht.

Der wesentliche Engpass bei der Bearbeitung von sehr Komplexen genetischen Informationen mit einem solchen Träger ist der Zugriff auf diese Information durch die begrenzte Zahl von Meßplätzen auf dem Träger. Ein solcher Meßplatz ist ein Reaktionsbereich, in dem bei der Herstellung des Träger DNA-Moleküle als spezifische Reaktionspartner, sog. Sonden, synthetisiert werden.

Für einen größeren Datendurchsatz gibt es prinzipiell zwei Möglichkeiten: Die eine besteht darin, die Anzahl der Meßplätze auf einem Reaktionsträger zu erhöhen. Die zweite beruht darauf, die Anzahl der unterschiedlichen Sonden zu steigern, die das System pro Zeit (und pro eingesetztem Geld) erzeugen und für Hybridisierung bereitstellen kann. Die zweite Möglichkeit hat etwas mit der Anzahl an Varianten zu tun, die im System generiert und für die Analyse zur Verfügung gestellt werden (Datendurchsatz).

Bei dem Begriff genetische Information muss unterschieden werden zwischen unbekannten Sequenzen, die zum ersten mal dekodiert werden (dies wird im allgemeinen unter dem Begriff Sequenzieren verstanden, auch *de novo* Sequenzierung) und bekannten Sequenzen, die aus anderen Gründen als dem erstmaligen Dekodieren identifiziert werden sollen. Solche anderen Gründe sind beispielsweise die Untersuchung der Expression von Genen oder die Verifizierung der Sequenz eines interessierenden DNA Abschnittes bei einem Individuum. Dies kann z.B. geschehen, um die individuelle Sequenz mit einem Standard zu vergleichen, wie bei der Mutationsanalyse von Krebszellen und der Typisierung von HIV Viren.

Für die *de novo* Sequenzierung werden bislang fast ausschließlich elektrophoretische Methoden verwendet. Am schnellsten ist die Kapillarelektrophorese.

Träger spielen für die *de novo* Sequenzierung bislang kaum eine Rolle. Dies liegt an prinzipiellen Limitationen: für den Informationsgewinn durch Sequenzvergleich müssen Sonden auf dem Träger bereitgestellt werden. Bei der Bearbeitung von unbekanntem Material braucht man sehr viele unterschiedliche Sonden (Varianten). Kein bislang bekanntes Verfahren ist in der Lage, die notwendigen Varianten-Zahlen für ein effektives Sequenzieren durch Sequenzvergleich von sehr großen DNA Mengen zu generieren. Solche sehr großen DNA Mengen liegen z.B. bei der Sequenzbestimmung von ganzen Genomen vor.

Bislang sind im wesentlichen zwei Verfahren zur Herstellung von Trägern bekannt. Beim ersten Herstellungsverfahren werden die fertigen Sonden einzeln entweder in einem Synthesizer (chemisch) oder aus isolierter DNA (enzymatisch) hergestellt und diese dann in Form winziger Tropfen auf die Oberfläche des Chips aufgebracht, und zwar jede einzelne Sorte an Sonden auf einen einzelnen Meßplatz. Das verbreitetste Verfahren hierzu leitet sich aus der Tintenstrahldrucktechnik ab, daher werden diese Verfahren unter dem Oberbegriff Spotting zusammengefaßt. Ebenfalls weit verbreitet sind Verfahren mit Nadeln. Nur durch die Mikro-Positionierung von Druckkopf oder Nadel kann später ein Signal auf dem Chip einer bestimmten Sonde zugeordnet werden (Array mit Zeilen und Spalten). Entsprechend genau müssen die Spotting-Geräte arbeiten.

Bei der zweiten Methode werden die DNA Sonden direkt auf dem Chip hergestellt, und zwar durch ortsspezifische Chemie (*in situ* Synthese). Dazu gibt es derzeit zwei Verfahren.

Das eine arbeitet mit den oben beschriebenen Spotting-Geräten, jedoch mit dem Unterschied, daß die winzigen Tropfen entsprechende Synthesechemikalien enthalten, so daß durch die Mikro-Positionierung dieser Chemikalien die ortsaufgelöste Chemie betrieben werden kann. Die Technologie erlaubt eine beliebige Programmierung der Sequenz der

entstehenden Sonden. Allerdings ist bisher der Durchsatz, das heißt die Anzahl der Sonden pro Zeit, nicht wirklich hoch genug, um große Mengen genetischer Information umzusetzen.

Sehr viel mehr Meßplätze pro Zeit lassen sich mit der zweiten Methode herstellen: die parallele Synthese der Sonden mit einer lichtabhängigen Chemie. Damit wurden bereits über 100.000 Meßplätze pro Chip in wenigen Stunden synthetisiert.

Das Verfahren wird mit zwei technischen Lösungen für die Belichtung betrieben. Die eine verwendet photolithographische Masken und erzeugt durch die hoch entwickelte Optik sehr viele Meßplätze auf dem DNA-Träger. Allerdings ist die Wahl der Sondensequenz sehr limitiert, da entsprechende Masken hergestellt werden müssen. Für das erfindungsgemäße Verfahren ist diese Herstellungsmethode daher wenig geeignet. Wesentlich aussichtsreicher sind Verfahren mit frei programmierbaren Sondensequenzendie auf Basis entsprechend steuerbarer Lichtquellen arbeiten. Solche Herstellungsverfahren für Sonden auf einem Träger sind u.a. in den Patentanmeldungen DE 198 39 254.0, DE 198 39 256.7, DE 199 07 080.6, DE 199 24 327.1, DE 199 40 749.5, PCT/EP99/06316 und PCT/EP99/06317 beschrieben.

Zusammenfassend läßt sich sagen, daß mit den bisher etablierten Techniken zur Bearbeitung größerer Mengen genetischer Information mit ganz oder teilweise unbekannter Zusammensetzung, nämlich Elektrophoreseverfahren und Biochip-Trägern, eine Limitation des Durchsatzes gegeben ist. Hochdurchsatzprojekte für die Neusequenzierung sind bisher auf Größensortierung mit Elektrophorese angewiesen (u.a. das Human Genom Projekt HUGO). Hier sind zwar Verbesserungen durch Miniaturisierung und Parallelisierung zu erwarten, aber keine Durchbrüche, da die Technik an sich nicht verändert werden kann. Elektrophorese kann die meisten Anwendungen von Biochips, wie z.B. Expressions-Muster oder Mutations-

Screening, nicht oder nur sehr viel langsamer leisten. Die bisher bekannten Biochips sind ihrerseits für Neusequenzierung ungeeignet, der Schwerpunkt liegt auf der hochparallelen Bearbeitung von Material auf Basis bekannter Sequenzen (u.a. in Form von synthetischen Oligonukleotiden als Sonden).

Beide Formate haben einen limitierten Durchsatz an genetischer Information. Um diesen Durchsatz zu erhöhen müssen neue Ansätze entwickelt werden. Das erfindungsgemäße Verfahren ist ein solcher Ansatz.

3. Gegenstand der Erfindung

Die Erfindung betrifft ein Verfahren zur Sequenzierung von Nukleinsäuren umfassend die Schritte:

(a) Durchführen eines ersten Hybridisierungszyklus umfassend

(i) Bereitstellen eines Trägers mit einer Oberfläche, die an einer Vielzahl von vorbestimmten Bereichen immobilisierte Hybridisierungssonden enthält, wobei die Hybridisierungssonden in einzelnen Bereichen jeweils eine unterschiedliche Basenfolge mit einer vorbestimmten Länge aufweisen,

(ii) Inkontaktbringen einer Probe, die zu sequenzierende Nukleinsäuren enthält, mit dem Träger unter Bedingungen, bei denen eine Hybridisierung zwischen den zu sequenzierenden Nukleinsäuren und dazu komplementären Sonden auf dem Träger erfolgen kann, und

(iii) Identifizieren der vorbestimmten Bereiche auf dem Träger, an denen eine Hybridisierung in Schritt (ii) erfolgt ist,

(b) Durchführen eines nachfolgenden Hybridisierungszyklus umfassend:

(i) Bereitstellen eines weiteren Trägers mit einer Oberfläche, die an eine Vielzahl von vorbestimmten Bereichen immobilisierte Hybridisierungssonden enthält, wobei die Hybridisierungssonden in einzelnen Bereichen jeweils eine unterschiedliche Basenfolge mit einer vorbestimmten Länge aufweisen, wobei

für den weiteren Träger Hybridisierungssonden mit einer Basenfolge ausgewählt werden, bei denen im vorhergehenden Zyklus eine Hybridisierung beobachtet worden ist, und wobei die ausgewählten Hybridisierungssonden um mindestens ein Nukleotid gegenüber einem vorhergehenden Zyklus verlängert werden,

(ii) Wiederholen von Schritt (a) (i) mit dem weiteren Träger, und

(iii) Wiederholen von Schritt (a) (iii) mit dem weiteren Träger, und

(c) gegebenenfalls Durchführen von weiteren nachfolgenden Hybridisierungszyklen jeweils mit Auswahl und Verlängerung der Hybridisierungssonden gemäß Schritt (b) (i), bis eine ausreichende Information über die zu sequenzierenden Nukleinsäuren vorliegt.

Das hier beschriebene Verfahren zur Sequenzierung von Nukleinsäuren durch Hybridisierung erlaubt mit Hilfe eines iterativen, dynamischen Aufbaus aller dafür notwendigen, spezifischen Sonden die Sequenzierung von Probenmaterial (auch viel größer 10 kBp) mit unbekannter Sequenz. Die Sequenzierung umfaßt sowohl eine Fragmentanalyse (einige Dutzend bis 100 Bp) als auch die Kartierung der Fragmente innerhalb der Ausgangssequenz.

Unter Träger oder Reaktionsträger sollen in diesem Zusammenhang sowohl offene als auch geschlossene Träger verstanden werden. Offene Träger können planar (z.B. Labordeckglas), aber auch speziell geformt (z.B. schalenförmig) sein. Bei allen offenen Trägern ist als Oberfläche eine Fläche auf der Außenseite des Trägers zu verstehen. Geschlossene Träger haben eine innenliegende Struktur, die beispielsweise Mikrokanäle, Reaktionsräume oder/und Kapillaren umfaßt. Hier sind als Oberflächen des Trägers die Oberflächen der zwei- oder dreidimensional ausgeprägten Mikrostruktur im Inneren des Trägers zu verstehen. Natürlich ist auch die Kombination von innenliegenden geschlossenen und außenliegenden offenen Oberflächen in einem Träger denkbar. Als Materialien für Träger kommen beispielweise Glas

wie Pyrax, Ubk7, B270, Foturan, Silizium und Siliziumderivate, Kunststoffe wie PVC, COC oder Teflon sowie Kalrez zum Einsatz.

Das in dem Verfahren benötigte Array muß nicht zwangsläufig auf einen Träger begränzt sein, es ist durchaus möglich ein "virtuelles Array" auf mehrere Träger zu verteilen. Bei Bedarf kann dadurch die Stellplatzanzahl vergrößert werden.

In einem geschlossenen System, das sowohl die Probenvorbereitung, die Fragmentierung und die Kartierung des Probenmaterials enthalten kann, siehe z.B. DE 199 24 327.1, DE 199 40 749.5 und PCT/EP99/06317, ergänzen und bedingen sich Datenerzeugung und Auswertung gegenseitig und bilden in ihrer Gesamtheit eine lernende Einheit. So werden z. B. mit Hilfe der ausgewerteten Daten eines Arrays neue Sondensequenzen bestimmt, die dann auf einem neuen Array synthetisiert werden. Dies erfolgt solange systematisch, bis die biologische Vielfalt, welche nur eine sehr geringen Teil der theoretisch Möglichen Variationen darstellt, schrittweise ganzheitlich erfaßt ist.

Bei dem erfindungsgemäßen Verfahren werden Sonden auf bzw. in dem Träger flexibel hergestellt, so daß ein Informationsfluß möglich wird. Jede neue Synthese des Arrays in aufeinanderfolgenden Zyklen kann die Ergebnisse eines vorangegangenen Experimentes berücksichtigen. Durch geeignete Wahl der Hybridisierungssonden, die Oligonukleotide, aber auch Nukleinsäureanaloga wie peptidische Nukleinsäuren sein können, in Bezug auf ihre Länge, Sequenz und Verteilung auf dem Reaktionsträger und durch eine Rückkopplung des Systems mit integrierter Signalauswertung wird ein effizientes Prozessieren von genetischer Information möglich.

Weiterhin betrifft die Erfindung einen Träger für die Sequenzierung von Nukleonsäuren mit einer Oberfläche, die an einer Vielzahl von vorbestimmten Bereichen immobilisierte Hybridisierungssonden enthält,

wobei die Hybridisierungssonden in einzelnen Bereichen jeweils eine unterschiedliche Basenfolge mit einer vorbestimmten Länge aufweisen, wobei die Hybridisierungssonden neben variablen Abschnitten einen oder mehrere für zumindest einen Teil der Sonden festgewählte Abschnitte aufweisen können.

Das Verfahren und der Träger können für die Sequenzbestimmung von Genomen, Chromosomen, Transkriptomen sowie zur Identifizierung von Polymorphismen in Nukleinsäuresequenzen, z.B. auf Ebene einzelner Individuen eingesetzt werden.

Die Bindung der Nukleinsäuren an Hybridisierungssonden an den jeweiligen Teilbereichen auf der Trägeroberfläche wird vorzugsweise über Markierungsgruppen nachgewiesen. Die Markierungsgruppen können dabei direkt oder indirekt an die zu sequenzierende Nukleinsäure gebunden werden. Vorzugsweise werden Markierungsgruppen verwendet, die optisch detektierbar sind, z.B. durch Fluoreszenz, Lichtbrechung, Lumineszenz oder Absorption. Bevorzugte Beispiele für Markierungsgruppen sind fluoreszierende Gruppen oder optisch nachweisbare Metallpartikel, z.B. Goldpartikel.

4. Ausführliche Beschreibung der Erfindung

4.1 (Zahlen-)Verhältnisse

Zu Beginn werden einige Verhältnisse erläutert, die im folgenden eine wichtige Rolle spielen:

In jeder, aus m Nukleotiden bestehenden Sequenz können maximal $m-n+1$ Teilsequenzen der Länge n auftreten. Dies bedeutet, daß für jede Gesamtsequenzlänge m eine spezifische Sequenzlänge n existiert, für die die Anzahl aller möglichen n-mere (4^n) die Anzahl m-n+1 der in der

Gesamtsequenz möglichen Teilsequenzen der Länge *n* überschreitet. Im menschlichen Genom z. B., das aus ca. $3,2 \times 10^9$ Nukleotiden besteht, können somit maximal ca. $3,2 \times 10^9$ Sequenzabschnitte einer beliebigen Länge *n* auftreten. Wählt man $n = 16$, so ist die Anzahl aller 16-mere mit 4^{16} deutlich größer als die maximale Anzahl der im menschlichen Genom auftretenden 16-mere. Es können also auf keinen Fall alle 16-mere und somit auch niemals alle längeren $(n+1)$-, $(n+2)$-mere, usw. im menschlichen Genom vorkommen.

Tabelle 1 zeigt das Verhältnis zwischen der Sequenzabschnittslänge *n*, der Sequenzlänge *m* und der in der Sequenz der Länge *m* enthaltenen maximalen Anzahl von Teilsequenzen der Länge *n*. In jeder Sequenz, die kürzer ist als der für *m* angegebene Wert, können nicht alle möglichen Abschnitte der angegebenen Länge *n* vorkommen.

Betrachtet man nun alle in einer Sequenz der Länge *m* auftretenden *n*-mere, die auf eine Teilsequenz der Länge *p* folgen, so ist die Anzahl dieser *n*-mere im Vergleich zu der oben beschriebenen Anzahl von m-n + 1 Teilsequenzen deutlich geringer.

Eine Sequenz, die alle 4^p möglichen *p*-mere enthält, muß eine minimale Länge von $k = 4^p + p1$ Nukleotiden aufweisen. Setzt man voraus, daß alle *p*-mere mit der gleichen Wahrscheinlichkeit vorkommen, so tritt in einer hinreichend lang gewählten Sequenz jedes *p*-mer im Mittel alle *k* Nukleotide einmal auf; in einer Sequenz der Länge *m* mit $m >> k$ also $l = m/k = m/4^p + p\text{-}1$ mal. Folglich können in einer solchen Sequenz mit Länge *m* auch maximal *l* *n*-mere beobachtet werden, die auf ein *p*-mer folgen.

Tabelle 1:

	Sequenzlänge	n-mere in der Sequenz
n	*m*	m - 4^n + 1
3	66	64
5	1028	1024
6	4101	4096
7	16390	16384
8	65543	65536
9	262152	262144
10	1048585	1048576
12	16777227	16777216
13	67108876	67108864
14	268435469	268435456
15	1073741838	1073741824
16	4294967311	4294967296
17	17179869200	17179869184
18	68719476753	68719476736
19	2,74878E+11	2,74878E+11
20	1,09951E+12	1,09951E+12
25	1,1259E+15	1,1259E+15

Wählt man z.B. im menschlichen Genom (einzelsträngig) ein beliebig aber fest gewähltes 3-mer und untersucht alle Sequenzabschnitte der Länge *n*, die auf dieses 3-mer folgen, findet man, bei einer vorausgesetzten Gleichverteilung aller *p*-mere, maximal 48.500.000 verschiedene *n*-mere.

Auch in diesem Fall gibt es eine charakteristische Grenze für die Vielfalt der Teilsequenzen. Wählt man die betrachteten Teilsequenzen länger als die der maximalen Vielfalt zugehörige Länge *n* , so gibt es mehr mögliche Varianten als in der untersuchten Sequenz vorkommen können. Beim menschlichen Genom (unter allen verallgemeinernden Voraussetzungen) ist dies eine Abschnittlänge von n = 13; insgesamt gibt es 4^{13} = 67108864 Sequenzen der Länge 13. Im menschlichen Genom können aber, wie oben errechnet, nur ca. 50.000.000 verschiedene Teilsequenzen nach einem frei gewählten 3-mer vorkommen. Für jede längere Teilsequenzlänge können auf keinen Fall alle möglichen Varianten im Genom vorkommen.

Tabelle 2 zeigt an einigen Beispielen den Zusammenhang zwischen der Sequenzlänge *m*, der Wahl von *p* und der Länge *n* der Teilsequenz, die nach dem *p*-mer betrachtet werden soll. In der dritten Spalte ist das unter idealisierten Annahmen durchschnittliche Vorkommen des gewählten *p*-mers in der Ausgangssequenz aufgetragen, daraus wird der Wert für *n* bestimmt, für den noch die komplette Vielfalt der *n*-mere nach dem *p*-mer vorkommen kann. Für jedes größer gewählte *p* oder für jede kürzer gewählte Sequenz trifft dies nicht mehr zu.

Ein längeres *p*-mer schränkt die Vielfalt innerhalb der untersuchten Sequenz deutlicher ein als ein kürzeres *p*-mer, da das längere *p*-mer im Verhältnis seltener auftritt.

Tabelle 2:

Sequenzlänge		Vorkommen d. p-mers		Anzahl *n*-mere
m	*p*	m/(4^p + p -1)	*n*	4^*n*
4352	2	256	4	256
16896	3	256	4	256
66304	4	256	4	256
17408	2	1024	5	1024
67584	3	1024	5	1024
265216	4	1024	5	1024
17825792	2	1048576	10	1048576
69206016	3	1048576	10	1048576
271581184	4	1048576	10	1048576
285212672	2	16777216	12	16777216
1107296256	3	16777216	12	16777216
4345298944	4	16777216	12	16777216
1140850688	2	67108864	13	67108864
4429185024	3	67108864	13	67108864
17381195776	4	67108864	13	67108864
4563402752	2	268435456	14	268435456
17716740096	3	268435456	14	268435456
69524783104	4	268435456	14	268435456

Das im folgenden beschriebene Verfahren macht sich diese Reduktion der Vielfalt zu Nutze. So ist es zum Beispiel nach den obigen Betrachtungen nicht notwendig, die komplette Menge aller 25-mere auf einem Array zu synthetisieren, wenn man eine Aussage darüber treffen will, welche 25-

mere in einer Probensequenz vorkommen. Je nach Länge der untersuchten Sequenz kann nur ein sehr geringer Anteil aller 25-mere in dieser Sequenz vorkommen, siehe Tabelle 1.

4.2 Dynamischer Arrayaufbau

Im Vergleich zu den bisher gängigen (statischen) Verfahren der Generierung von Trägerchips, ist es erfindungsgemäß möglich, schnell von einem Array zum nachfolgenden Array zu lernen und dadurch ein Vielfaches der bisherigen Informationsmenge zu erhalten.

Können in kurzer Zeit verschiedene Arrays unter Verwendung der, nach Auswertung des Vorgängerarrays, erhaltenen Informationen erzeugt werden, so wird das System zu einem "lernenden" System. Mit dieser Methode können die oben erwähnten 25-mere einer Sequenz bestimmt werden, ohne sie in ihrer Vielfalt (4^{25} = 1.125899907 x 10^{15}) synthetisieren zu müssen.

Man kann beispielsweise mit einer variablen Sondenlänge *s* beginnen, mit der die mögliche Vielfalt (4^s) aller *s*-mere auf dem Array synthetisierbar ist. Falls alle möglichen 4^s Sequenzvariationen nicht auf einem einzigen Träger erzeugt werden können, ist es möglich auch eine begrenzte Anzahl von mehreren Trägern für einen Hybridisierungszyklus zu verwenden. Liegt die Länge der Sonden unter dem in Tabelle 1 ermittelten Wert *n*, so ist es möglich, daß alle auf dem Array erzeugten Sequenzen in der Ausgangssequenz vorkommen, wahrscheinlich ist es aber nicht. Zudem nimmt diese Wahrscheinlichkeit mit wachsender Länge der Sonden ab. Auf jeden Fall können aber nicht mehr als die in Tabelle 1 errechneten Teilsequenzen in der Sequenz vorkommen.

Im nächsten Schritt werden alle Sonden, die auf dem Vorgängerarray ein Signal erzeugt haben, auf einem neuen Array synthetisiert und um jeweils

mindestens ein Nukleotid an allen möglichen Variationen verlängert, d.h. bei einer Verlängerung um ein Nukleotid entstehen vier unterschiedlich verlängerte Hybridisierungssonden. Spätestens ab der in Tabelle 1 dargestellten Teilsequenzlänge *n* wird sich die Anzahl der Signale nicht mehr vergrößern, weil ihre Anzahl (unter idealisierten Annahmen) nicht größer sein kann als die maximale Anzahl der unterschiedlichen Teilsequenzen in der Ausgangssequenz. Unter "normalen" Voraussetzungen wird es Signale geben, die nach idealisierten Voraussetzungen nicht hätten entstehen dürfen. Diese Sonden können zunächst weiter aufgebaut werden, durch verlängerte Sonden und die dadurch resultierenden spezifischeren Bindungen können mögliche Fehler im Laufe der Iteration eliminiert werden. In der Praxis wird zudem nie die komplette Vielfalt aller möglichen Teilsequenzen in einer zu untersuchenden Sequenz auftreten, so daß deutlich weniger Signale als die maximal mögliche Anzahl erzeugt werden.

Je nach Anzahl der Stellplätze und der Länge der zu untersuchenden Sequenz ist es bevorzugt, die Sondenlänge des ersten Arrays so zu wählen, daß nach der Hybridisierung von maximal 25% aller Stellplätze Signale ausgehen. Durch dieses Vorgehen wird gewährleistet, daß die Anzahl der Sonden im nächsten Schritt nicht zunimmt. Die Sonden auf dem neuen Array können somit um eine Base länger als die Sonden auf dem Vorgängerarray gewählt werden, ohne daß sich die Anzahl der Sonden vergrößert.

Die Länge *m* der Sequenz (in diesem Fall ein Einzelstrang, für einen Doppelstrang gilt ähnliches) muß für eine solche Wahl der Startsonden kleiner sein als die erlaubte Anzahl der Signale, in Formeln: $m \leq 4^{s-1} + s-1$, wobei *s* die Sondenlänge ist. Auf einem Array mit Sondenlänge $s = 6$ kann also eine Sequenz der maximalen Länge $m = 4^5 + 5 = 1029$ bearbeitet werden, so daß nach der Hybridisierung auf jeden Fall von weniger, bzw. von maximal 25% aller Sonden Signale ausgehen. Die folgende Tabelle 3

zeigt die bevorzugte Länge s der Startsonden in Abhängigkeit von der Länge m der zu bestimmenden Sequenz.

Tabelle 3:

Sondenlänge *s*	Sequenzlänge *m*
5	260
6	1029
7	4102
8	16391
9	65544
10	262153
11	1048586
12	4194315
13	16777228
14	67108877
15	268435470
16	1073741839
17	4294967312
20	2,74878E+11
22	4,39805E+12
25	2,81475E+14

Da in einer Sequenz der Länge *m* Teilsequenzen der Länge *s* durchaus mehrfach auftreten können, reduziert sich die rechnerische Anzahl von m-s + 1 Teilsequenzen der Länge s oftmals in der Praxis. In einem solchen Fall ist eine kleinere Sondenlänge ausreichend. Da die Anzahl sich wiederholender Sequenzen zu Beginn aber nicht bekannt ist, ist der oben bestimmte Wert als oberer Grenzwert anzusehen. Die Anzahl der Signale wird durch wiederholte Auftreten einer Teilsequenz reduziert, aber niemals vergrößert.

Einige Zahlenbeispiele:

Für das menschliche Genom mit $3,2 \times 10^9$ Nukleotiden pro Strang ist eine Sondenlänge von 17 Basen ausreichend, um theoretisch sicher zu stellen, daß an weniger als 25% aller Stellplätze auf dem Array eine Bindung stattfindet. Für *E.coli* mit 4 639 221 Nukleotiden sind bereits Sonden der

Länge 13 ausreichend. Die Stellplatzanzahl aller folgenden Arrays wird die Anzahl der Stellplätze auf diesen Arrays nicht überschreiten.

Wählt man die Länge der Sonden auf dem ersten Array nicht nach der oben beschriebenen Methode, so pendelt sich die Anzahl der Signale auf jeden Fall im Laufe des Verfahrens unter den maximalen Wert von m-n + 1 ein, wobei *n* die im ersten Abschnitt beschriebene Länge ist, für die die Vielfalt aller *n*-mere größer ist als die Anzahl der in der Ausgangssequenz möglichen *n*-mere. Wählt man zu Beginn eine zu kurze Sondenlänge, so wird sich die Anzahl der benötigten Stellplätze in den nächsten Schritten zunächst bis zu maximal 4^{n-1} Stellplätzen erhöhen und dann stagnieren. Wählt man die Sonden zu lang, so werden bei der Hybridisierung deutlich weniger als 25% aller Stellplätze erfolgreich sein, so daß sich die Anzahl der benötigten Stellplätze im nächsten Schritt automatisch reduziert.

Wie im ersten Abschnitt beschrieben, läßt sich die Vielfalt der Teilsequenzen in einer Sequenz der Länge *m* noch weiter reduzieren, indem man nur Sequenzabschnitte betrachtet, die auf eine vorher festgelegte Abfolge von Nukleotiden folgt. Auch in diesem Fall läßt sich die Länge der Sonden auf dem ersten Array wie oben bestimmen. Für ein Array, auf dem alle Kombinationen der Länge s = n + p synthetisiert werden, die mit dem *p*-mer beginnen oder enden, bedeutet dies, daß nur von maximal 25% (d.h. $1/4^n$%) 4^{n-1} aller Stellplätze auf diesem Array Signale ausgehen dürfen. Somit kann auf einem Array mit Sondenlänge s = n + p und einem beliebigen, aber für alle oder einen Teil der Sonden festgewählten Abschnitt der Länge p eine Sequenz der Länge *m* mit $m \leq 4^{n-1} \times (4^p + p - 1)$ hybridisiert werden, ohne daß die theoretisch mögliche Anzahl der Stellplätze, von denen Signale ausgehen können, 25% aller Stellplätze überschreitet wobei; *n* ist dabei der im ersten Abschnitt berechnete Wert ist.

Das Verhältnis zwischen der maximalen Länge der Ausgangssequenz und der Länge der Sonde, sowie der *p*-mere ist in Tabelle 4 für einige Beispiele

dargestellt. Für das menschliche Genom genügt bei einem festgewählten 3-mer eine Sondenlänge von $n + p = 17$ Nukleotiden, um die erlaubte Anzahl der Stellplätze, die ein Signal liefern, nicht zu überschreiten. Die Anzahl der zu synthetisierenden Sonden ist in jedem Fall 4^n, also die Menge aller Möglichkeiten, den flexiblen Sondenteil aufzubauen.

Die oben, sowie die im ersten Abschnitt berechneten Werte gelten für eine Gleichverteilung der betrachteten p-mere. In den meisten Sequenzen gilt diese idealisierte Annahme nicht, es treten unter Umständen stark unterschiedliche Verteilungen der einzelnen Nukleotide auf. Kennt man daher z. B: bei DNA- / RNA-Sequenzen den A-T -, bzw. C-G- Gehalt der zu untersuchenden Sequenz, so lassen sich Wahrscheinlichkeiten für die einzelnen *p*-mere berechnen. Durch eine Gewichtung bei der Berechnung der maximalen Sequenzlänge mit Hilfe der Wahrscheinlichkeit für das Auftreten des gewählten *p*-mers werden sich in einigen Fällen die in den Tabellen 2 und 4 aufgeführten Werte verschieben.

Tabelle 4: Maximal mögliche Länge der Ausgangssequenz im Verhältnis zur Sondenlänge und ihrer Zusammensetzung.

		Sondenlänge	Sequenzlänge
n	*p*	*s = p + n*	*m*
4	3	7	4224
4	4	8	16576
5	3	8	16896
5	4	9	66304
8	3	11	1081344
8	4	12	4243456
10	3	13	17301504
10	4	14	67895296
12	3	15	276824064
12	4	16	1086324736
14	3	17	4429185024
14	4	18	17381195776
15	3	18	17716740096

Somit bietet der dynamische Aufbau einer Folge von Arrays den Vorteil, daß nach Auswertung der Informationen des bzw. der Vorgänger–Arrays ein

neues Array aufgebaut werden kann, das die benötigten Daten liefert. Es ist möglich, Kenntnis über Teilsequenzen in der Ausgangssequenz von spezifischer Länge, z.B. von 25 Basen und mehr, zu gewinnen, ohne alle möglichen Kombinationen dieser Länge aufbauen zu müssen. Das Verfahren pendelt sich automatisch auf eine maximale Signalanzahl und somit auf eine maximale Stellplatzanzahl pro Array ein.

Im folgenden wird eine Anwendung beschrieben, die sich mit dem oben beschriebenen dynamischen Arrayaufbau realisieren läßt.

4.3 Dynamisches Sequenzieren durch Hybridisierung (DSBH)

An dieser Stelle wird zunächst das allgemeine Prinzip des DSBH beschrieben, das im wesentlichen durch einen flexiblen Aufbau der Arrays möglich wird; im nächsten Abschnitt folgen mögliche Umsetzungen dieses Prinzips.

Wie oben beschrieben, kommen *p*-mere in einer zu bestimmenden Sequenz mit unterschiedlichen Wahrscheinlichkeiten vor, die sich z. B. bei DNA-Sequenzen durch Kenntnis des A-T und G-C Gehalts der Sequenz bestimmen lassen. Der Grundgedanke des DSBH besteht nun darin, *p*-mere auszuwählen, die in regelmäßigen Abständen in der Sequenz vorkommen, sie lassen sich als "Inseln" auffassen, deren Sequenz bereits bekannt ist. Von diesen festgewählten Orten bekannter Sequenz (Points of Known Sequence, kurz POKS) ausgehend, wird nun die Probensequenz bestimmt. Dazu werden zunächst drei Arten von Sonden auf den Arrays benötigt:

(1) Sonden mit festgewählten Sequenzen am 3´- Ende,

(2) Sonden mit festgewählten Sequenzen am 5´-Ende,

(3) Sonden mit festgewählten Sequenzen im Innern, z.B. im Zentrum der Sequenz.

Die Sonden (1), (2) und (3) können gemeinsam oder/und nacheinander auf dem gleichen Träger oder auf unterschielidhcen Trägern eingesetzt werden. Für die beiden ersten Sondentypen werden alle Kombinationen einer vorgegebenen Länge synthetisiert, wobei die Gegensequenz zum gewählten POKS einmal am 3´-Ende der Sequenz und einmal am 5´-Ende der Sequenz aufgebaut wird. Durch die Hybridisierung der Ausgangssequenz gegen die Sonden dieses Arrays erhält man dann Informationen über alle Nukleotidkombinationen der vorgegebenen Länge einmal in 3´-5´-Richtung zum POKS hin und einmal in 3´-5´-Richtung vom POKS weg. Nach dem oben beschriebenen Vorgehen zum dynamischen Aufbau der Arrays werden alle Sonden der Stellplätze, die ein Signal erzeugt haben, auf einem neuen Array synthetisiert und dabei jeweils um ein Nukleotid in allen vier Variationen verlängert. Bei einer hinreichend großen Anzahl von Stellplätzen auf dem Array können auch zwei oder mehr Iterationsschritte auf einem Array abgearbeitet werden, d.h. es kann eine Verlängerung um zwei oder mehr Nukleotide erfolgen.

Bei der Verlängerung der Sonden ist zu beachten, daß Sonden, bei denen die zum POKS komplementäre Sequenz am 3´-Ende aufgebaut wird, in 5´-Richtung verlängert werden, und Sonden mit der komplementären POKS-Sequenz am 5´-Ende entsprechend in 3´-Richtung. Hat die Iteration eine maximale Sondenlänge erreicht, so ist zu beiden Seiten jedes POKS die Abfolge der Nukleotide auf der Länge der maximalen Sondenlänge bekannt. Die Sondenlänge wird dabei entweder durch die Möglichkeiten des verwendeten Systems beschränkt oder durch einen Kompromiß aus der benötigten Zeit bis zum endgültigen Ergebnis und dessen Genauigkeit.

Mit Hilfe der dritten Sondenart wird der Zusammenhang zwischen den oben bestimmten Sequenzen hergestellt. Es werden nun all die Sondensequenzen bestimmt, die die POKS-Gegensequenz im Zentrum haben und davor, bzw. dahinter Teile der durch die ersten beiden Sonden gewonnen Informationen. Diese Sonden werden auf einem neuen Array aufgebaut; nach der

Hybridisierung und Auswertung der Signale sind alle Möglichkeiten bekannt, zu denen die durch die ersten zwei Sondenarten bestimmten Sequenzen zusammengesetzt werden dürfen.

Diese Information kann man genauso durch einen iterativen Arrayaufbau erhalten, bei dem alle Kombinationen einer bestimmten Länge vor und nach der POKS-Gegensequenz aufgebaut werden. Nach Auswertung der Signale werden die relevanten Sonden wie oben beschrieben weiter verlängert, jetzt in beide Richtungen, usw. Bei einer hinreichend großen Stellplatzanzahl kann man diese Iterationsschritte aber durch den sofortigen Aufbau der benötigten Sonden zur maximalen Länge vermeiden.

Das Array mit der dritten Sondenart löst hochparallel eine kombinatorische Aufgabe, die ohne einen flexiblen Arrayaufbau nur mit sehr großem Rechenaufwand mit Hilfe von Computern lösbar ist. Die Verlagerung dieser Aufgabe auf das Array bedeutet einen erheblichen Zeitgewinn gegenüber einer Kombinatorik am Rechner und liefert zudem verläßlichere Daten.

Werden die POKS nun entsprechend gewählt, so kann mit der oben beschriebenen Methode die Ausgangssequenz wieder zusammengesetzt werden, indem die Überlappungen der durch die einzelnen POKS bestimmten Teilsequenzen verglichen und kombiniert werden.

In den folgenden Punkten 5 und 6 sind nun zwei besonders bevorzugte Ausführungsformen des erfindungsgemäßen Verfahrens im Detail erläutert.

5. Dynamische Sequenzierung durch Hybridisierung (DSBH) mit statistisch gewählten festen Sondenabschnitten (POKS)

5.1 Voraussetzungen

Das Verfahren zur Sequenzierung mit statistisch, bzw. durch das Verfahren gewählten POKS, sowie die zugehörige Probenvorbereitung werden für einen Einzelstrang beschrieben. Mit dem gleichen Verfahren ist auch die Sequenzierung doppelsträngiger Nukleinsäuren möglich.

5.1.1 Probenvorbereitung

Die hier beschriebene Sequenzierung geht von einzelsträngigen Nukleinsäuren aus. Diese können im einfachsten Fall direkt in Form einzelsträngiger RNA oder DNA aus Viren, Bakterien, Pflanzen, Tieren oder dem Menschen isoliert werden. In der Mehrzahl der Fälle werden die einzelsträngigen Nukleinsäuren aber ausgehend von dsDNA durch spezielle *in vitro* Verfahren erzeugt. Hierzu zählen z.B. asymmetrische PCR (erzeugt ssDNA), PCR mit derivatisierten Primern, die eine selektive Hydrolyse eines einzelnen Stranges im PCR-Produkt ermöglichen, oder die Transkription durch RNA-Polymerasen (erzeugt ssRNA). Als Matrize kann bei der Transkription neben nicht klonierter einzelsträngiger DNA vor allem auch in spezielle Vektoren (z.B. Plasmidvektoren mit einem Promotor; Plasmidvektoren mit zwei unterschiedlich orientierten Promotoren für eine bestimmte oder zwei unterschiedliche RNA-Polymerasen) klonierte dsDNA eingesetzt werden. Die in die Plasmide klonierte Insert-DNA oder die bei der PCR eingesetzte DNA-Matrize können zum einen aus Viren, Bakterien, Pflanzen, Tieren oder dem Menschen isoliert werden, zum anderen aber auch *in vitro* durch reverse Transkription, RNaseH-Behandlung und anschließende Amplifikation (z.B. durch PCR) aus ssRNA erzeugt werden. Als RNA-Matrizen können rRNAs, tRNAs, mRNAs und snRNAs sowie *in*

vitro erzeugte Transkripte (entstanden z.B. durch Transkription mit SP6-, T3- oder T7-RNA-Polymerase) eingesetzt werden.

Die für die Sequenzierung vorgesehenen, einzelsträngigen Nukleinsäuren werden sequenzspezifisch oder/und sequenzunspezifisch fragmentiert (z.B. durch sequenz(un)spezifische Enzyme, Ultraschall oder Scherkräfte), wobei eine im wesentlichen homogene Längenverteilung der Bruchstücke/Hydrolyseprodukte angestrebt wird. Wird keine homogene Längenverteilung der Fragmente erreicht, kann anschließend eine Längenfraktionierung durch gelelektrophoretische und/oder chromatographische Verfahren durchgeführt werden.

Die entstandenen Fragmente können mit Markierungsgruppen, z.B. fluoreszierenden Agenzien oder radioaktiven Isotopen markiert werden. Die Markierung erfolgt dabei bevorzugt an den Enden der Fragmente (terminale Markierung). 3'-terminale Markierungen können unter Verwendung geeigneter Synthone z.B. mit der terminalen Transferase oder der T4 RNA-Ligase durchgeführt werden. Werden für die Fragmentierung *in vitro* erzeugte RNA-Transkripte eingesetzt, kann die Markierung auch vor der Fragmentierung durch bei der Transkription eingesetzte markierte Nukleotide erfolgen (interne Markierung).

Die markierten, fragmentierten Nukleinsäuren können dann in einer geeigneten Hybridisierungslösung gegen den mit einem Sondenarray beschichteten Träger hybridisiert werden.

5.2 Auswahl der festgelegten Sondenabschnitte (POKS)

In der folgenden Variante des Verfahrens zur Sequenzierung mit POKS dienen nach unterschiedlichen Kriterien ausgewählte *p*-mere als POKS; sie können zu verschiedene Zeitpunkten des Verfahrens bestimmt werden.

Zum einen kann zu Beginn des Verfahrens eine festgelegte Anzahl POKS bestimmt werden. Hier bietet es sich an, die Kombinationen (*p*-mere) auszuwählen, die in der Ausgangssequenz mit der höchsten Wahrscheinlichkeit vorkommen. Dies ist möglich, da die einzelnen Nukleotide und somit auch die einzelnen *p*-mere wie im ersten Abschnitt beschrieben mit unterschiedlichen Wahrscheinlichkeiten in der Probensequenz vorkommen. Kennt man z. B. bei DNA-Sequenzen den G-C, bzw. A-T Gehalt dieser Sequenz, so können also diejenigen *p*-mere bestimmt werden, die am wahrscheinlichsten, und somit am häufigsten in der Sequenz auftreten. Es sind ebenso andere Methoden zu einer Wahl der POKS zu Beginn des Verfahrens denkbar, z.B. aus Erfahrungswerten oder durch eine willkürliche Bestimmung.

Zum anderen kann es sinnvoll sein, nur wenige, bzw. einen POKS zu Beginn des Verfahrens festzulegen und alle folgenden POKS aus den bis dahin gewonnen Sequenzinformationen zu bestimmen. Durch dieses Vorgehen lernt das Verfahren aus den bisher generierten Daten und bestimmt, welche Daten für den weiteren Verlauf des Verfahrens und das Zusammensetzen der Informationen wichtig sind. Die ersten POKS müssen nicht notwendiger Weise vom Anwender vorgegeben werden, sie können z. B. wie oben erläutert vom System durch Bestimmung der Wahrscheinlichkeiten für die potentiellen POKS, aus Erfahrungswerten oder willkürlich bestimmt werden.

Bei einer Wahl der POKS zu Beginn des Verfahrens muß zunächst die Anzahl der POKS festgelegt werde. Diese kann z. B. aus Erfahrungswerten bestimmt, oder statistisch berechnet werden, indem sie so groß gewählt wird, daß der Abstand zwischen zwei POKS rein rechnerisch deutlich kleiner ist als die vorgegebene maximale Sondenlänge auf den Arrays.

Werden die POKS erst im Laufe des Verfahrens bestimmt, so kann ihre Anzahl entweder vorher festgelegt werden, s.o., so daß das Verfahren mit dem Erreichen der maximalen POKS-Anzahl abbricht, oder es werden so

lange weitere POKS bestimmt, bis andere Abbruchkriterien erfüllt sind. Zum Beispiel kann das Verfahren abgebrochen werden, wenn eine Sequenz von einer vorgegebenen Länge zusammengesetzt wurde, die alle Ansprüche an eine potentielle Lösung des Problems erfüllt. Ebenso kann das Verfahren z. B. dann beendet werden, wenn sich sie bisher zusammengesetzten Sequenzen an keinem der beiden Enden weiter verlängern lassen.

5.3 Vorgehensweise

Das Verfahren beruht im wesentlichen auf dem oben beschriebenen dynamischen Arrayaufbau, da dieser es erlaubt, Sequenzinformationen von spezifischer Länge zu erhalten, ohne dazu alle Sonden in ihrer Vielfalt erzeugen zu müssen. Außerdem wird die parallele "Rechenleistung" der Arrays genutzt, die zeit- und rechenaufwendige Vorgänge im Computer überflüssig macht.

5.3.1 Verschiedene Sondentypen auf dem Array

Für alle zu Beginn festgelegten POKS werden die drei oben beschriebenen Sondentypen auf einem oder mehreren Arrays synthetisiert, d.h. einmal werden alle Kombinationen einer vorgegebenen Länge mit der POKS-Gegensequenz am 3´-Ende und einmal mit dieser Sequenz am 5´-Ende erzeugt. Durch die Hybridisierung mit der Ausgangssequenz erhält man nach der Signalauswertung Informationen in (ungefährer) Sondenlänge über die Paarungen der Nukleotide rechts und links von diesen POKS. Mit Hilfe der Signale können wie oben beschrieben iterativ neue Sonden erzeugt werden. Dies wiederholt sich, bis eine maximale Sondenlänge erreicht ist. Zu diesem Zeitpunkt kennt man in der Ausgangssequenz alle möglichen Kombinationen auf maximaler Sondenlänge zu beiden Seiten jedes POKS.

Tabelle 5:

	N	**P**	N	5'-Ende
	N	**P**	N	
5	N	**P**	N	
	N	N	N	
	N	N	N	
	N	N	N	
	N	N	**P**	
10	N	N	**P**	
	N	N	**P**	
	N	N	N	
	N	N	N	
	N	N	N	
15	**P**	N	N	
	P	N	N	
	P	N	N	3'-Ende

Tabelle 5 zeigt die drei verschiedenen Sondentypen mit den POKS (PPP) bzw. deren komplementärer Sequenz am 3'-Ende, am 5'-Ende und im Inneren der Sonde

Mit Hilfe des dritten Sondentyps wird nun der Zusammenhang zwischen diesen Informationen geklärt. Jede Sonde enthält nun im Zentrum die Gegensequenz zu den gewählten POKS, zu beiden Seiten dieser Sequenz werden nun in verschiedenen Sonden alle möglichen Kombinationen einer bestimmten Länge erzeugt. Durch das gleiche iterative Vorgehen wie bei den beiden ersten Sondentypen gewinnt man Informationen über alle Kombinationen der bisher erkannten Sequenzen, die in der Ausgangssequenz auftreten. Wenn die sich aus der Anzahl aller möglichen Kombinationen der erkannten Sequenzen ergebende Zahl der benötigten Stellplätze für den dritten Sondentyp geringer ist als die Stellplatzanzahl auf dem Array, können die Teile der erkannten Sonden des 1. und 2. Typs direkt in die neuen Sonden übernommen werden. Eine Iteration ist in diesem Fall nicht notwendig. Für die direkte Erzeugung aller möglichen Zusammenhänge zwischen den erkannten Sequenzen werden deutlich weniger Stellplätze benötigt.

5.3.2 Zusammensetzen der ersten Sequenzinformationen

Nach der Auswertung der Arrays mit Sonden des dritten Typs und einem Zwischenschritt im Rechner sind alle Kombinationen der Länge

$$k = 2 \times \text{Maximale Sondenlänge} - \text{POKS-Länge}$$

bekannt, die in der Ausgangssequenz auftreten können; sie haben alle einen POKS in der Mitte der Sequenz.

Mit Hilfe der POKS lassen sich diese Teilsequenzen nun erweitern. Dazu wird in jeder Teilsequenz zu einer oder beiden Seiten des mittleren POKS eine neue Stelle gesucht, an der einer der verwendeten POKS auftritt. Wird ein POKS gefunden, so vergleicht man die Sequenzinformation zu beiden Seiten dieses POKS mit allen Teilsequenzen, die genau diesen POKS enthalten. Dieses Vorgehen ermöglicht die Verknüpfung der einzelnen Teilsequenzen, es entsteht ein Baum aller Varianten, in denen diese Sequenzen kombinierbar sind.

Die folgende Tabelle 6 zeigt die Überschneidung zweier Teilsequenzen in einer DNA Sequenz, die mit Hilfe eines POKS erkannt wurde.

Tabelle 6:

ATGGAGCACTTGG**PPP**CCTACG**PPP**GTCA

TTGG**PPP**CCTACG**PPP**GTCATTGGCAGTA

In der oberen Sequenz von Tabelle 6 wurde ein weiterer POKS an Position 7 rechts nach dem POKS in der Mitte gefunden. Der Vergleich mit der zweiten Sequenz, die den "erkannten" POKS in der Mitte der Sequenz hat, hat ergeben, daß eine größtmögliche Überschneidung zwischen den beiden

Sequenzen besteht, und zwar von Position eins der zweiten Sequenz bis zu Position 20 dieser Sequenz.

Wurden alle POKS bereits zu Beginn des Verfahrens bestimmt, so sind nun alle möglichen Nachbarschaftsverhältnisse der Teilsequenzen bekannt. Die Nukleotidkombinationen können zur Gesamtsequenz zusammengesetzt werden, dazu wird der Baum aller Kombinationsmöglichkeiten durchlaufen und sinnvoll erscheinende Teilsequenzen werden zu einer Gesamtsequenz vereint. Falls repetitive Teilsequenzen auftreten, wird der Algorithmus nach einigen Zyklen abgebrochen; ein mögliches Abbruchkriterium ist dabei zum Beispiel die angenommene Länge der Ausgangssequenz.

Alle potentiellen Lösungssequenzen müssen zum Schluß noch auf ihre Richtigkeit überprüft werden, damit der Fehler zwischen der bestimmten Lösungssequenz und der Ausgangssequenz möglichst gering ist.

5.3.3 Bestimmung neuer POKS

Wurden nicht alle POKS gleich zu Beginn des Verfahrens festgelegt, so ist es nun möglich, neue POKS aus den bereits bekannten Sequenzteilen zu bestimmen. Hierzu gibt es mehrere Varianten. Zum einen können alle Teilsequenzen zu einer Seite der POKS in der Mitte jeder Sequenz auf die am häufigsten auftretenden p-mere untersucht werden, wobei p die Länge der zu wählend POKS ist, die entweder vorher festgelegt oder im Verfahren optimiert werden kann. Durch diese Wahl der POKS kann im nächsten Schritt für eine Mehrzahl, bzw. für alle bis jetzt bekannten Teilsequenzen eine Sequenz bestimmt werden, durch die sich die bisher detektierten Sequenzen verlängern lassen. Um sicher zu stellen, daß für jede Teilsequenz eine Folgesequenz, bzw. eine Vorgängersequenz gefunden wird, werden eventuell relativ viele POKS benötigt. Mit den neu bestimmten POKS werden die gleichen Sonden erzeugt wie mit den zu Beginn gewählten POKS. Mit den dadurch gewonnenen Informationen ergeben sich neue Möglichkeiten,

die bekannten Teilsequenzen zusammenzusetzen und zu verlängern. Sollten die Abbruchkriterien des Verfahrens noch nicht erfüllt sein, so werden aus den neu bestimmten Sequenzen wiederum POKS bestimmt und mit deren Hilfe neue Informationen gewonnen.

Um die Anzahl der benötigten POKS zu verringern, ist es sinnvoll, die mit den zu Beginn des Verfahrens gewählten POKS gewonnenen Informationen zunächst zu längeren Sequenzen zusammenzusetzen. Diese längeren Sequenzen werden, falls erforderlich, untereinander verglichen und kürzere Sequenzen, die auch in längeren Sequenzen zu finden sind, gestrichen. Die restlichen Sequenzen enden alle auf Teilsequenzen für die kein Nachfolger bestimmt werden kann, bzw. beginnen alle mit Sequenzen, für die es keinen Vorgänger gibt. In diesen "Endsequenzen" werden nun wie oben *p*-mere bestimmt, die häufig vorkommen. Die *p*-mere dienen als neue POKS, für die wieder die drei Sondentypen erzeugt werden und somit nach der Signalauswertung alle möglichen Basenkombinationen um die POKS bekannt sind.

Nur in der Anfangssequenz und der Endsequenz der zu untersuchenden Sequenz können POKS gefunden werden, ohne daß sich diese Sequenzen weiter verlängern lassen. Werden diese Teilsequenzen im Verfahren erkannt, so werden sie gesondert behandelt und nicht in die Bestimmung neuer POKS einbezogen.

Aufgrund der Wahl der neuen POKS überschneiden sich die neu bestimmten Sequenzen nun zum Teil mit den bereits bekannten längeren Sequenzen, diese werden nun, soweit möglich, in beide Richtungen verlängert. Zudem werden alle Kombinationen erzeugt, die durch die neuen POKS entstehen und noch nicht in den bisher bekannten Sequenzen enthalten sind. Aus den neuen "Endsequenzen" werden wieder neue POKS erzeugt; dies geschieht so lange, bis eines der Abbruchkriterien erfüllt wird.

Neben den oben aufgeführten Methoden zur Bestimmung der POKS sind natürlich auch andere Vorgehensweisen denkbar, bei denen POKS nach den einzelnen Teilschritten des Verfahrens bestimmt werden. Unter anderem kann sich eine Kombination aus verschiedenen Methoden als sinnvoll erweisen.

Durch die selbständige Wahl der neuen POKS entwickelt sich im System ein Lernprozeß, bei dem sich die Auswertung der Daten und die Zusammensetzung neuer Arrays zur Gewinnung neuer Daten gegenseitig bedingen.

5.3.4 Endgültiges Zusammensetzen und Verifizierung der Sequenzen

Bestimmt man die POKS zu Beginn des Verfahrens, so werden die erkannten Teilsequenzen in allen möglichen Kombinationen zu langen Sequenzen zusammengesetzt. Bei einer entsprechenden Auswahl der POKS überlappt jede Teilsequenz mit einer anderen, so daß sich die Ursprungssequenz unter den kombinierten Möglichkeiten befindet. Um herauszufinden, welche der Sequenzen diejenige ist, die das Problem am besten löst, werden zunächst alle Sequenzen untereinander auf Überlappungen überprüft. Treten solche Überlappungen auf, und überschreitet eine aus den sich überlappenden Teilsequenzen zusammengesetzte Sequenz nicht die geschätzte oder bekannte Länge der Probensequenz, so werden die Sequenzen weiter kombiniert. Kurze Sequenzen, die komplett in längeren Sequenzen enthalten sind, werden gestrichen.

Neben der Sequenzlänge ist der Vergleich mit allen auf den Arrays detektierten Teilsequenzen ein Anhaltspunkt, um die Sequenz zu bestimmen, die mit der Probensequenz am besten übereinstimmt. In der Lösungssequenz sind im Idealfall alle, zumindest aber ein großer Teil der auf den Arrays mit den ersten beiden Sondentypen bestimmten Sequenzen

enthalten, auf keiner Fall dürfen vor oder nach einem POKS Basenkombinationen auftreten, die nicht auf den Arrays erkannt wurden.

Ist zudem eine Quantifizierung der erhaltenen Signale möglich, kann also zumindest annähernd bestimmt werden, wie oft eine detektierte Sequenz in der Ursprungssequenz vorkommt, so ist dies ein weiteres Kriterium während der Verifizierung; es darf keine Sequenz häufiger als erkannt vorkommen.

Außer den oben aufgeführten Kriterien ist es natürlich möglich, die gleiche Sequenz zur Kontrolle mit anderen POKS zu untersuchen und die Ergebnisse zu vergleichen, ein Prozeß, der bei einer hohen Stellplatzdichte auf den Arrays durchaus parallel verlaufen kann.

Werden die POKS erst im Verlauf des Verfahrens bestimmt, so kann schon in jedem Schritt überprüft werden, ob die einzelnen Sequenzen nur Teilsequenzen enthalten, die auch in der Probensequenz vorkommen, oder ob Sequenzen auftreten, die nicht auftreten dürfen und eine Sequenz damit Lösungssequenz ausscheidet. Genauso kann (bei der oben angesprochenen Quantifizierung der Signale) schon nach jedem Schritt sichergestellt werden, daß eine Teilsequenz nur so oft eingebunden wird wie es zulässig ist.

5.3.5 Abbruchkriterien

Bei einer vorher festgelegten Anzahl von POKS kann das Verfahren automatisch abgebrochen werden, wenn nach bzw. bei der Bestimmung neuer POKS diese Anzahl überschritten wird, bzw. wenn bei vorgegebenen POKS alle dadurch erhaltenen Informationen verarbeitet wurden.

Sind sowohl die POKS als auch deren Anzahl frei wählbar, so muß ein anderes Abbruchkriterium gefunden werden. Zunächst ist die Bestimmung von p-meren natürlich begrenzt durch deren Anzahl, da es genau 4^p p-mere

gibt. Je nach Wahl von p ist diese Anzahl relativ hoch und damit zu groß, um als natürliches Abbruchkriterium zu dienen.

Ohne jedes Vorwissen über die Beschaffenheit der zu untersuchenden Sequenz (z.B. ohne Kenntnis ihrer Länge) kann das Verfahren dann abgebrochen werden, wenn für jede theoretisch verlängerbare, erkannte Teilsequenz ein Nachfolger, bzw. ein Vorgänger gefunden wurde. Zu diesem Zeitpunkt liegt die komplette Sequenzinformation der Ausgangssequenz vor, so daß durch eine erneute Bestimmung von POKS keine neuen Informationen gewonnen werden können.

Ist die Länge der zu untersuchenden Sequenz bekannt, so kann die zyklische POKS-Bestimmung beendet werden, sobald eine Sequenz gefunden wurde, deren Länge mit der ungefähren Ausgangslänge übereinstimmt, und die (fast) alle auf den Arrays erkannten Teilsequenzen enthält.

Zudem können für die zusammengesetzten Sequenzen während des Verfahrens Wahrscheinlichkeiten für ihre "Richtigkeit", bzw. Werte zur Fehlerabschätzung bestimmt werden, so daß das Verfahren abbrechen kann, sobald ein vorher gesetzter Schwellenwert für den Fehler unterschritten wird.

5.3.6 Wiederholungen innerhalb der Ausgangssequenz und repetitive Sequenzen

Treten in der Probensequenz Wiederholungen auf, so kann es in dem oben beschriebenen Baum aller möglichen Sequenzkombinationen zu einem Ringschluß kommen, der das Zusammensetzen der Sequenzen erschwert.

Dabei ist die Länge der sich wiederholenden Sequenzabschnitte von wesentlicher Bedeutung. Wiederholungen, die kürzer sind als die maximale

Sondenlänge (bei Verwendung aller 3 Sondentypen), bzw. kürzer als die halbe maximale Sondenlänge bei ausschließlicher Verwendung des 3. Sondentyps, stellen kein Problem beim Zusammensetzen dar. Treten Wiederholungen auf, die länger sind als die oben beschriebenen, die aber kürzer als die Gesamtlänge der Teilsequenzen minus Länge der POKS, so können diese durch geschicktes Verschieben der POKS, d.h. durch die Wahl eines neuen POKS, der sehr nahe am POKS im Zentrum der Sequenz liegt, aufgelöst werden. Treten längere Wiederholungen auf, so wird nach ihrem Auftreten der Algorithmus zum Zusammensetzen abgebrochen, dadurch entstehen mehrere Teilsequenzen von unterschiedlicher Länge, die jeweils um die Länge der Wiederholungen überlappen. Durch den Einsatz anderer Verfahren, wie z.B. PCR, oder der Wahl neuer Sondentypen kann der Zusammenhang zwischen diesen Teilsequenzen geklärt werden.

Ein möglicher weiterer Ansatz zur Lösung der durch Wiederholungen bedingten Phänomene ist die Kenntnis über die ungefähre Länge der Ausgangssequenz. Wird bei dem Versuch, die erkannten Teilsequenzen zusammenzusetzen, diese Länge deutlich überschritten, so wurden vermutlich Teilsequenzen zu häufig eingebaut. Eine solche Sequenz kann nicht als Ergebnis des Verfahrens zugelassen werden.

Ist es darüber hinaus möglich, durch eine Quantifizierung der nach der Hybridisierung erhaltenen Signale eine Größenordnung für die Häufigkeit des Auftretens jeder Sonde in der Ausgangssequenz festzulegen, so wird die Länge der Ausgangssequenz nicht unbedingt als Abbruchkriterium benötigt.

Auch für den Fall, daß in der Probensequenz repetitive Teile auftreten, d.h. nicht unterbrochene Wiederholungen relativ kurzer Sequenzen, erleichtert die mögliche Quantifizierung der Signale auf den Arrays das Zusammensetzen der Sequenz.

5.4 Sequenzieren mit langen Sonden

Ist es möglich, die Sondenlängen in dem oben beschriebenen Verfahren hinreichend groß zu wählen, so kann auf den Aufbau der ersten beiden Sondentypten für jeden POKS verzichtet werden. Die Sonden können dann so lang gewählt werden, daß die Wahrscheinlichkeit, für einen weiteren POKS in ihrer Sequenz groß genug ist, um Überlappungen zu garantieren. Wie oben beschrieben werden für den nun ausschließlich relevanten 3. Sondentyp, der die Gegensequenz der gewählten POKS in der Mitte der Sequenz enthält, alle Kombinationen einer vorgegebenen Länge erzeugt, gegen diese wird hybridisiert und signalliefernde Sonden werden im nächsten Schritt weiter aufgebaut. Dabei ist es möglich, jede Sonde gleich in beide Richtungen vom POKS weg zu verlängern, oder abwechselnd in die eine und dann in die andere, bis die maximal mögliche Länge erreicht wird. Je nach Anzahl der Stellplätze können wieder mehrere Iterationsschritte auf einem Array abgearbeitet werden.

Die Verwendung von langen Sonden macht unter Umständen den Aufbau der ersten beiden Sondentypen überflüssig. Dies bedeutet eine Reduktion der Stellplätze und somit der benötigten Arrays. Zum anderen können eventuelle Fehler, die durch die rechnerische Verlängerung der Sonden des dritten Typs mit Hilfe der Sonden des ersten und zweiten Typs entstehen, ausgeschlossen werden.

6. Dynamische Sequenzierung durch Hybridisierung (DSBH) mit durch Enzym-Erkennungsstellen gewählten festen Abschnitten (POKS)

Eine weitere Variante des Verfahrens ist die Integration der POKS bereits in die Probenvorbereitung, indem mittels sequenzspezifischen Nukleasen das Probenmaterial in entsprechende Fragmente geschnitten wird. Als POKS dienen dann automatisch die Basen, die die Nuklease-Erkennungssequenzen bilden.

6.1.1 Probenvorbereitung

Die Probenvorbereitung für diese Variante des Verfahrens geht zunächst von dsDNA aus. Diese dsDNA kann zum einen als genomische, chromosomale DNA, als extrachromosomales Element (z.B. als Plasmid) oder als Bestandteil von Zellorganellen aus Viren, Bakterien, Tieren, Pflanzen oder dem Menschen isoliert werden, zum anderen aber prinzipiell auch *in vitro* durch reverse Transkription, RNaseH-Behandlung und anschließende Amplifikation (z.B. durch PCR) aus ssRNA erzeugt werden. Als RNA-Matrizen können neben rRNAs, tRNAs, mRNAs und snRNAs auch *in vitro* erzeugte Transkripte (entstanden z.B. durch Transkription mit SP6-, T3- oder T7-RNA-Polymerase) eingesetzt werden.

Die isolierte oder *in vitro* synthetisierte dsDNA wird dann mit einer Restriktionsendonuklease oder mit einem Gemisch aus mehreren Restriktionsendonukleasen hydrolysiert, wobei doppelsträngige Subfragmente mit definierten Anfangs- und/oder Endsequenzen entstehen. Anzahl und Länge der entstehenden Subfragmente können durch die Auswahl geeigneter Enzyme (dies können auch durch Proteindesign veränderte oder erzeugte Enzyme sein) gesteuert werden. Zur Längenfraktionierung können der Hydrolyse gelelektrophoretische und/oder chromatographische Trennprozesse folgen. Für die Erzeugung von RNA-Subfragmenten können Ribozyme eingesetzt werden.

Die erzeugten Subfragmente werden vorzugsweise nach der Fraktionierung markiert. Obwohl die Markierung prinzipiell auch vor der Denaturierung möglich ist (z.B. durch das Auffüllen 3'-kohäsiver Enden mit einer DNA-Polymerase), werden die Subfragmente bevorzugt nach der Denaturierung, also auf der Ebene einzelsträngiger Subfragmente, markiert. Die Markierung erfolgt vorzugsweise mittels fluoreszierender Agenzien (z.B. Fluorescein oder Cy5), möglich sind aber auch andere Markierungsverfahren wie z.B. der Einbau radioaktiver Isotope. Die Markierungsgruppen werden

hauptsächlich in Form markierter Nukleotid-Derivate an die Subfragmente gekoppelt. Die Kopplung am 3'-Terminus kann z.B. durch die T4-RNA-Ligase oder durch die terminale Transferase (unter Verwendung entsprechender Nukleotid-Derivate) erfolgen.

Die markierten, einzelsträngigen Subfragmente können dann in einer geeigneten Hybridisierungslösung gegen den mit einem Sondenarray beschichteten Träger hybridisiert werden.

6.2 Verfahrensablauf

Die in geeigneter Weise aufbereitete Probe wird durch ein Schnittenzym in möglichst kleine Subfragmente zerlegt. Die komplementäre Sequenz zur Nukleotidabfolge des Schnittenzyms bildet hierbei direkt die POKS Sequenz, das bedeutet, die möglichen POKS werden durch die zur Verfügung stehenden Enzyme vorgegeben. Das statistische Verhalten der Fragmentlänge und -anzahl ist analog zu den frei gewählten POKS bedingt durch die Ausgangssequenz und die verwendete Schnittsequenz.

Die so enzymatisch zerkleinerte Probe wird nach der Länge der Subfragmente sortiert, d.h. fraktioniert. Markierte Subfragmente, welche nicht länger als die maximale Sondenlänge sind, werden zur Analyse, gemäß beschriebenen Verfahren, auf den Array gegeben. Die Sonden, welche beim ersten Array einen Hybridisierungspartner unter den Subfragmenten in der Probe gefunden haben, werden entsprechend zyklisch bis zur maximalen Sondenlänge verlängert. Dadurch werden alle Subfragmente der Ausgangsprobe bezüglich ihrer Nukleotidabfolge bestimmt.

Die längeren Subfragmente werden einem weiteren Probenvorbereitungszyklus zugeführt. Dabei kann es sich wiederum um eine enzymatische Fragmentierung, aber auch ein geeignetes Amplifikationsverfahren oder das

vorher beschriebene rein statistische POKS Verfahren und die zugehörige Probenvorbereitung handeln.

Bei Bedarf können auch mehrere Enzym POKS gleichzeitig in der Probenvorbereitung und in der anschließenden zyklischen Arrayanalyse eingesetzt werden. Diese Subfragmente können durch die enzymatische POKS Sequenz am Anfang bzw. Ende der Sonden einwandfrei zugeordnet und parallel verfolgt werden.

Für den Aufbau der Sonden ergeben sich in dieser Variante des DSBH-Verfahrens durch die Vorgabe der Enzymsequenzen zwei Möglichkeiten. Zum einen kann die komplette Sequenz an den Enden der Sonden aufgebaut werden, zum anderen kann es genügen, nur den Teil der Enzymsequenz nach dem Schnittpunkt zu synthetisieren. Tabelle 7 stellt die beiden Möglichkeiten am Beispiel einer DNA-Sequenz dar, in der die Sequenz des Enzyms Alu I (AGCT) auftritt. Die Schnittstelle dieses Enzyms liegt zwischen dem zweiten und dritten Nukleotid.

Tabelle 7

5´-Ende	NNNNNNNNNNNNN AG ¦ CT NNNNNNNNNNNNN	3´-Ende
3´-Ende	NNNNNNNNNNNNN TC ¦ CA NNNNNNNNNNNNN	5´-Ende

Nach der Hydrolyse und der Denaturierung in der Probenvorbereitung erhält man in diesem Fall vier Fragmente. Zwei von ihnen beginnen, in 5´-3´ Richtung gelesen, mit den Nukleotiden CT, die beiden anderen Enden auf AG. Um die in beiden Richtungen auf die Enzymsequenz folgenden Nukleotide erkennen zu können, müssen auf dem Array nun die drei oben beschriebenen Sondentypen synthetisiert werden, siehe Tabelle 8.

Im linken Teil der Tabelle 8 wird die komplette Enzymsequenz als POKS verwendet, der Aufbau erfolgt völlig analog zur Methode mit statistisch

gewählten POKS. Für den Aufbau der im rechten Teil dargestellten Sonden wird die Enzymsequenz an ihrem Schnittpunkt in zwei Teile zerlegt. Um die im obigen Sequenzbeispiel mit den Nukleotiden CT beginnenden Fragmente detektieren zu können, werden Sonden mit dem den Nukleotiden GA am 3´-Ende erzeugt, um die beiden anderen Fragmente bestimmen zu können, werden alle Sonden einer vorgegbenen Länge erzeugt, die die Nukleotide TC am 5´-Ende tragen. Das Hybridisierungsverhalten auf dem Array muß für beide Sondentypen gleich sein. Im linken Fall fungieren die Nukleotide TC als eine Art Linker.

Für die jeweils dritte Sondenarte muß die Probe anders vorbereitet werden. Entweder wird die zu untersuchende Sequenz statistisch, z.B. mit Ultraschall zerlegt, oder z. B. mit einem Enzym geschnitten, dessen Sequenz keiner der zur Probenvorbereitung verwendete Enzymsequenzen entspricht.

Tabelle 8:

N	**A**	N	5'-Ende	N	**C**	N	5'-Ende
N	**G**	N		N	**T**	N	
N	**C**	N		N	N	N	
N	**T**	N		N	N	N	
N	N	N		N	N	N	
N	N	N		N	N	N	
N	N	**A**		N	N	**A**	
N	N	**G**		N	N	**G**	
N	N	**C**		N	N	**C**	
N	N	**T**		N	N	**T**	
N	N	N		N	N	N	
A	N	N		N	N	N	
G	N	N		N	N	N	
C	N	N		**A**	N	N	
T	N	N	3'-Ende	**G**	N	N	3'-Ende

Das Zusammensetzen der einzelnen detektierten Fragmente zu einer Gesamtsequenz erfolgt analog zur beschriebenen Variante mit statistisch gewählten POKS.

Der wesentliche Vorteil der Erzeugung der POKS in der Probenvorbereitung durch Schnittenzyme ist ein niedrigerer Bedarf an Probenmaterial. Durch die enzymatische Zerlegung der Ausgangssequenz entstehen nur Subfragmente mit der POKS Sequenz am Ende. Bei einer Ausgangssequenz mit beispielsweise 3.000 Basen und einer mittleren Subfragmentlänge von 60 Basen entstehen ca. 500 Subfragmente. Beim Zerlegen der gleichen Ausgangssequenz in alle möglichen Subfragmente für die frei wählbaren POKS (aber mit der gleichen Nukleotidsequenz wie das Enzym sie aufweist) entstehen entsprechend 3.000 – 60 + 1 = 2.941 Subfragmente von denen nur 500 die POKS Sequenz am Ende aufweisen. Im Vergleich wird für die Enzym POKS also nur 500 / 2.941 = 0.17 entsprechend 17% des Probenmaterials benötigt.

Die wesentlichen Nachteile der enzymatischen POKS sind die notwendige Entwicklung der geeigneten Schnittenzyme, die geringe Flexibilität und der höherer Aufwand in der Probenvorbereitung. Die Entwicklung der entsprechenden Enzyme zum Beispiel mittels Proteindesign ist arbeitsaufwendig. Die Bereitstellung in der Probenvorbereitung erhöht den logistischen Aufwand im System. Außerdem muß eine zyklische Probenvorbereitung mit einer integrierten Längenfraktionierung etabliert werden. Diese ist notwendig um die längeren Subfragmente abzutrennen und weiter zu zerkleinern.

Beide Ansätze (frei wählbare und enzymatische POKS) lassen sich auch kombinieren. So könnten statistisch sehr erfolgreiche POKS als Enzyme in der Probenvorbereitung bereitgestellt werden. Sind diese Enzym POKS verbraucht wird entsprechend mehr amplifiziert und die frei wählbaren POKS eingesetzt.

7.1.1 Freigewählte POKS mit allen 3 Sondentypen

In diesem Beispiel wird die Sequenzierung einer 3060 Nukleotide langen einzelsträngigen Teilsequenz aus dem *E.coli* Genom mit Hilfe verschiedener POKS von drei Nukleotiden Länge simuliert. Die während der Simulation erzeugten Daten sind Idealdaten, die mögliche Fehler, wie z. B. möglichen Abbruch während der Synthese oder Probleme bei der Signalauswertung noch nicht berücksichtigen.

Mit Hilfe der durch die Simulation des Arrayaufbaus, der Hybridisierung und der Signalauswertung erzeugten Daten läßt sich die Ausgangssequenz wieder in ihrer Gesamtheit zusammensetzen.

Zu Beginn des Verfahrens wird der A-T-, G-C- Gehalt der Sequenz bestimmt. Daraufhin wird der POKS mit der höchsten Wahrscheinlichkeit, in diesem Fall GCG, als Start-POKS gewählt. Mit diesem POKS wird die Synthese der Sonden auf dem ersten Array simuliert. Dazu werden alle drei Sondentypen mit der Gegensequenz zum POKS an den oben näher beschriebenen Positionen in den Sonden erzeugt. Der variable Anteil der Sonden hat in diesem Beispiel eine Länge von 5 Nukleotiden, für jeden Sondentyp werden also Stellplätze benötigt, also insgesamt 3072. Um eine eventuell deutlich größere Anzahl von Stellplätzen auszunutzen, kann es sinnvoll sein, gleich zu Beginn längere Sonden zu synthetisieren.

Nach der Hybridisierung gehen von jeweils 82 Stellplätzen, deren Sonden die POKS-Gegensequenz an ihren Enden haben und von 81 Stellplätzen, deren Sonden die POKS-Sequenz in der Mitte haben, Signale aus. Auf dem nächsten Array werden also insgesamt 980 (82 x 4 + 81 x 4 + 81 x 4) Stellplätze benötigt, um für jeden signalgebenden Stellplatz vier neue Stellplätze mit jeweils um eine Base verlängerten Sonden aufbauen zu können.

An dieser Stelle ist es möglich, gleich mehrere Iterationsschritte auf einem Array abzuarbeiten, wenn die Anzahl der vorhandenen Stellplätze hinreichend groß ist. Dazu kann jede relevante Sonde auf dem neuen Array um zwei, drei oder mehr Nukleotide erweitert werden. Bei einer Verlängerung um zwei Nukleotide werden pro Stellplatz dann 16 neue Stellplätze benötigt, bei einer Verlängerung um drei Nukleotide entsprechend 64 Stellplätze, bei 4 Nukleotiden 256 Stellplätze, usw. In der Simulation, in der die Stellplatzanzahl eine untergeordnete Rolle spielt, wird für jeden Iterationsschritt ein neues Array erzeugt.

Die Sondenlänge von insgesamt 5 + 3 = 8 Nukleotiden ist in diesem Fall bereits so spezifisch lang, daß sich die Anzahl der benötigten Stellplätze in keinem der folgenden Iterationsschritte deutlich vergrößert, sie pendelt sich nach ungefähr 3 Schritten auf 340 Stellplätze pro Sondentyp, also insgesamt auf 1020 Stellplätze ein.

Insgesamt werden die Sonden bis zu einer Länge von 25 Nukleotiden aufgebaut, so daß nach der Auswertung des letzten Arrays alle in der Ausgangssequenz auftretenden 22-mere nach und vor dem ersten POKS bekannt sind. Mit Hilfe des dritten Sondentyps werden alle möglichen Zusammenhänge zwischen diesen Teilsequenzen bestimmt, diese Sequenzen können rechnerisch mit den Sequenzen des ersten und zweiten Sondentyps auf jeweils 47 Nukleotide verlängert werden.

Es ist mit dem dynamischen Arrayaufbau somit gelungen, alle 22-mere nach und vor dem POKS zu bestimmen, ohne alle 22-mere (4^{22} = 1,759218604 x 10^{13}) erzeugen zu müssen.

Im nächsten Schritt wird in den jetzt bekannten zusammengesetzten Teilsequenzen mit dem POKS in der Mitte die POKS-Sequenz rechts und links dieses POKS gesucht. Wird die POKS-Sequenz ein zweites Mal in einer Teilsequenz gefunden, so wird der entsprechende Abschnitt mit allen

Teilsequenzen verglichen, die den POKS in der Mitte haben. Da alle Sequenzen um den POKS nun bekannt sind, muß es eine Sequenz geben, mit der es eine Überschneidung gibt. Nach dem ersten POKS gelingt es bereits, die erkannten Teilsequenzen zu längeren Sequenzen bis zu 248 Nukleotiden Länge zusammenzusetzen. Durch Auswertung der Enden dieser Sequenzen werden zwei neue POKS (CTG, GAA) bestimmt, einer für jedes Ende, mit denen nun wieder Arrays aufgebaut werden. Wie oben wird mit einer variablen Länge von 5 Nukleotiden begonnen, die bis zu einer Länge von 22 Nukleotiden gesteigert wird. Die Anzahl der benötigten Stellplätze pendelt sich nach wenigen Zyklen auf 312 pro Sondentyp ein, so daß pro Iterationsschritt insgesamt 936 x 2 Stellplätze benötigt werden.

Wie gehabt werden in den detektierten Sequenzen die POKS-Sequenzen gesucht und diese Sequenzen gegebenenfalls verlängert. Nach den ersten drei POKS können Sequenzteile bis zu einer Länge von 456 Nukleotiden zusammengesetzt werden. Um die Sequenz in der vollen Länge erkennen und zusammensetzen zu können werden noch vier weitere POKS (GCC, CAG, TCA, ATC) benötigt, die aus den bisher ausgewerteten Daten und einem weiteren Zyklus bestimmt werden. Die Anzahl der in den letzen beiden Zyklen (Arrayaufbau, Hybridisierung, iterative Verlängerung der Sonden bis zu 25 Nukleotiden) benötigten Stellplätze pro Iterationsschritt liegt bei 200 bis 370 Stellplätzen pro Sondentyp. Nach dem letzen Zyklus kann die Ausgangssequenz komplett zusammengesetzt werden.

Die Array-Größe und die Anzahl der nach jedem Schritt gewählten POKS ist in diesem Beispiel nicht optimiert worden. Es ist möglich, daß eine größere Anzahl von POKS zu Beginn des Verfahrens die Anzahl der benötigten Stellplätze / Arrays reduzieren würde. Zudem erscheint es sinnvoll, auf jedem Array mehrere Iterationsschritte auf einmal abzuarbeiten, um die Anzahl der verfügbaren Stellplätzen auszunutzen. Geht man in diesem Beispiel von einer Array-Größe von 400.000 Stellplätzen aus, und optimiert das Verfahren, so können auf dem ersten Array Sonden mit einem variablen

Teil von 8 Nukleotiden aufgebaut, also mit einer Gesamtlänge von 11 Nukleotiden. Damit werden die vorhandenen Stellplätze allerdings erst zur Hälfte ausgenutzt, was eine Wahl von zwei POKS zu Beginn sinnvoll erscheinen läßt.

Auch bei einer Ausgangslänge von 11 Nukleotiden pro Sonden gehen nur von ca. 85 Stellplätzen pro Sondentyp Signale aus, so daß auf dem nächsten Array insgesamt 1020 Stellplätze aufgebaut werden müssen. Somit können auf diesem Array 5 Iterationsschritte abgearbeitet werden, dazu werden 261.124 Stellplätze benötigt. Mit zwei weiteren Arrays, auf denen wiederum jeweils 1024 Sonden pro signalgebenden Stellplatz des Vorgängerarrays aufgebaut werden können, lassen sich die relevanten Sonden auf jeweils 25 Nukleotide verlängern. Für den ersten POKS werden somit 4 Arrays benötigt; dabei sind die einzelnen Arrays noch nicht ideal ausgelastet.

Um in den nächsten Schritten zwei POKS auf einmal untersuchen zu können, muß die Anzahl der Iterationsschritte pro Array auf vier reduziert werden, so daß für jedes POKS-Paar insgesamt vier bis fünf Arrays benötigt werden, insgesamt, inklusive der Arrays für den ersten POKS, also 16 bis 19 Arrays.

Bei Beispielen mit längeren Sequenzen ist zu beobachten, daß die Anzahl der benötigten POKS nicht notwendigerweise mit der Länge der Sequenz wächst, vielmehr gelingt es z. B. verschiedene Sequenzen von 20.000 Nukleotiden Länge mit 9 bis 11 POKS zusammenzusetzen. Das Verfahren wird somit für längere Sequenzen immer rentabler.

8. Anwendungen

Das erfindungsgemäße Verfahren ermöglicht die systematische Sequenzanalyse von teilweise oder gänzlich unbekannten Nukleinsäuren in einer Probe.

In einer Ausführungsform werden mithilfe des Verfahrens Genome ganz oder teilweise sequenziert. Die Teile können durch Auswahl und Isolierung einzelner Chromosomen, durch Klonieren genomischer DNA (z.B. in *Bacterial Artificial Chromosomes* BAC oder *Yeast Artificial Chromosomes* YAC) oder durch andere Verfahren generiert werden.

In einer anderen Ausführungsform werden cDNA-Populationen, die z.B. aus einer klonierten Bibliothek oder direkt aus einer isolierten mRNA hergestellt sein können, ganz oder zum Teil sequenziert. Im Ergebnis handelt es sich dann um eine Transkriptom-Sequenzierung. Dies kann bei gleichzeitiger Bearbeitung unterschiedlcher Proben aus unterschiedlichen Quellen, z.B. Zellen in unterschiedlichem Zustand, so geschehen, daß in einer Variante nur solche Sequenzen weiterverfolgt werden, die unterschiedlich sind, in einer anderen nur solche, die gleich sind.

In einer Ausführungsform kann es von Interesse sein, daß sog. Polymorphismen, z.B. Einzelnukleotid-Polymorphismen, identifiziert oder für die Auswahl der POKS verwendet werden.

Weiterhin kann das erfindungsgemäße Sequenzierungsverfahren für diagnostische Zwecke, beispielsweise für eine individualisierte oder mehrstufige Diagnostik eingesetzt werden. Das Verfahren eignet sich auch zur Entwicklung einer individualisierten, patientenabhängigen Medikamentierung bzw. zur patientenabhängigen Entwicklung oder/und Modifizierung von pharmazeutischen Substanzen. Das Verfahren kann in Verbindung mit einem Netzwerk oder/und einer Datenbank zu einer

dezentralen patientennahen Analyse und Identifizierung von Krankheitsbildern bzw. Krankheitserregern und deren Mutationen eingesetzt werden. Außerdem ist das Verfahren zur molekularen Diagnostik sowie zur vergleichenden Genomik geeignet, z.B. zum Einsatz in der Forschung, zur Aufklärung der Funktionalität von einzelnen Genen oder Genomen von Organismen. Das Verfahren kann weiterhin zur Mutationsanalyse, z.B. unter anderem zur Untersuchung des Einflusses von beispielsweise Umwelteinflüssen, Medikamenten, Strahlung oder/und Giften von Organismen eingesetzt werden.

Ansprüche

1. Verfahren zur Sequenzierung von Nukleinsäuren umfassend die Schritte:

(a) Durchführen eines ersten Hybridisierungszyklus umfassend

(i) Bereitstellen eines Trägers mit einer Oberfläche, die an einer Vielzahl von vorbestimmten Bereichen immobilisierte Hybridisierungssonden enthält, wobei die Hybridisierungssonden in einzelnen Bereichen jeweils eine unterschiedliche Basenfolge mit einer vorbestimmten Länge aufweisen,

(ii) Inkontaktbringen einer Probe, die zu sequenzierende Nukleinsäuren enthält, mit dem Träger unter Bedingungen, bei denen eine Hybridisierung zwischen den zu sequenzierenden Nukleinsäuren und dazu komplementären Sonden auf dem Träger erfolgen kann, und

(iii) Identifizieren der vorbestimmten Bereiche auf dem Träger, an denen eine Hybridisierung in Schritt (ii) erfolgt ist,

(b) Durchführen eines nachfolgenden Hybridisierungszyklus umfassend:

(i) Bereitstellen eines weiteren Trägers mit einer Oberfläche, die an eine Vielzahl von vorbestimmten Bereichen immobilisierte Hybridisierungssonden enthält, wobei die Hybridisierungssonden in einzelnen Bereichen jeweils eine unterschiedliche Basenfolge mit einer vorbestimmten Länge aufweisen, wobei für den weiteren Träger Hybridisierungssonden mit einer Basenfolge ausgewählt werden, bei denen in einem vorhergehenden Zyklus eine Hybridisierung beobachtet

worden ist, und wobei die ausgewählten Hybridisierungssonden um mindestens ein Nukleotid gegenüber einem vorhergehenden Zyklus verlängert werden,

(ii) Wiederholen von Schritt (a) (i) mit dem weiteren Träger, und

(iii) Wiederholen von Schritt (a) (iii) mit dem weiteren Träger, und

(c) gegebenenfalls Durchführen von weiteren nachfolgenden Hybridisierungszyklen jeweils mit Auswahl und Verlängerung und Auswahl der Hybridisierungssonden gemäß Schritt (b) (i), bis eine ausreichende Information über die zu sequenzierenden Nukleinsäuren vorliegt.

2. Verfahren nach Anspruch 1,
dadurch gekennzeichnet,
daß die zu sequenzierenden Nukleinsäuren aus doppelsträngiger DNA, einzelsträngiger DNA und RNA ausgewählt werden.

3. Verfahren nach Anspruch 1 oder 2,
dadurch gekennzeichnet,
daß die zu sequenzierenden Nukleinsäuren vor dem Inkontaktbringen mit dem Träger fragmentiert werden.

4. Verfahren nach Anspruch 3,
dadurch gekennzeichnet,
daß durch die Fragmentierung und gegebenenfalls eine nachfolgende Längenfraktionierung Nukleinsäurefragmente mit einer vorbestimmten, z.B. im wesentlichen homogenen Längenverteilung erzeugt werden.

5. Verfahren nach Anspruch 3 oder 4,
dadurch gekennzeichnet,
daß die Fragmentierung sequenzunspezifisch erfolgt.

6. Verfahren nach Anspruch 3 oder 4,
dadurch gekennzeichnet,
daß die Fragmentierung sequenzspezifisch erfolgt.

7. Verfahren nach einem der vorhergehenden Ansprüche,
dadurch gekennzeichnet,
daß die zu sequenzierenden Nukleinsäuren Markierungsgruppen, insbesondere optisch detektierbare Markierungsgruppen wie Fluoreszenz- oder Metallpartikelmarkierungen tragen.

8. Verfahren nach Anspruch 7,
dadurch gekennzeichnet,
daß direkte oder indirekte Markierungen verwendet werden.

9. Verfahren nach einem der vorhergehenden Ansprüche,
dadurch gekennzeichnet,
daß im ersten Hybridisierungszyklus Sonden mit einer Länge s ausgewählt werden und alle möglichen 4^s Sequenzvariationen an den vorbestimmten Bereichen des Trägers erzeugt werden.

10. Verfahren nach einem der vorhergehenden Ansprüche,
dadurch gekennzeichnet,
daß im ersten Hybridisierungszyklus Sonden mit einer Länge s ausgewählt werden, so daß nach Inkontaktbringen mit der Probe an maximal 25% der vorbestimmten Bereiche eine Hybridisierung mit den zu sequenzierenden Nukleinsäuren erfolgt.

11. Verfahren nach einem der vorhergehenden Ansprüche,
dadurch gekennzeichnet,
daß im ersten Hybridisierungszyklus Sonden mit einer Länge s so ausgewählt werden, daß sie mit der Länge m der zu bestimmenden Sequenz in folgender Beziehung stehen:

$$m \leq 4^{s-1} + s - 1$$

12. Verfahren nach einem der vorhergehenden Ansprüche,
dadurch gekennzeichnet,
daß in einem oder mehreren Hybridisierungszyklen Sonden verwendet werden, die neben variablen Abschnitten der Länge n einen oder mehrere für zumindest einen Teil der Sonden festgewählte Abschnitte der Länge p aufweisen.

13. Verfahren nach Anspruch 12,
dadurch gekennzeichnet,
daß im ersten Hybridisierungszyklus die Länge n des variablen Sondenanteils so gewählt wird, daß alle möglichen 4^n Sequenzvariationen an den vorbestimmten Bereichen des Trägers erzeugt werden.

14. Verfahren nach Anspruch 12 oder 13,
dadurch gekennzeichnet,
daß die Länge p des festgewählten Abschnitts und die Länge n der variablen Abschnitte so ausgewählt werden, daß sie mit der Länge m der zu bestimmenden Sequenz in folgender Beziehung stehen:

$$m \leq 4^{n-1} (4^p + p - 1)$$

15. Verfahren nach einem der Ansprüche 12 bis 14,
dadurch gekennzeichnet,
daß die Länge der festgewählten Abschnitte p 2, 3, oder 4 Nukleotide beträgt.

16. Verfahren nach einem der Ansprüche 12 bis 15,
dadurch gekennzeichnet,
daß Sonden verwendet werden ausgewählt aus (1) Sonden mit den festgewählten Abschnitten p am 3'-Ende, (2) Sonden mit festgewählten Abschnitten p am 5'-Ende und (3) Sonden mit festgewählten Abschnitten p im Inneren der Sequenz.

17. Verfahren nach Anspruch 16,
dadurch gekennzeichnet,
daß Sonden mit festgewählten Abschnitten p im Inneren der Sequenz verwendet werden.

18. Verfahren nach Anspruch 16 oder 17,
dadurch gekennzeichnet,
daß die Sonden (1), (2) und (3) gemeinsam oder/und nacheinander auf dem gleichen Träger oder auf unterschiedlichen Trägern eingesetzt werden.

19. Verfahren nach einem der Ansprüche 12 bis 18,
dadurch gekennzeichnet,
daß die festgewählten Abschnitte p zu Beginn des Verfahrens oder/und aufgrund der Resultate von vorhergehenden Hybridisierungszyklen festgelegt werden.

20. Verfahren nach einem der Ansprüche 12 bis 19,
dadurch gekennzeichnet,
daß die festgewählten Abschnitte willkürlich, aufgrund statistischer oder/und aufgrund biochemischer Überlegungen bestimmt werden.

21. Verfahren nach einem der Ansprüche 12 bis 20,
dadurch gekennzeichnet,
daß die festgewählten Abschnitte aufgrund der Basenfolge von Enzym- oder/und Ribozym-Erkennungssequenzen, z.B. von Nukleasen bestimmt werden.

22. Verfahren nach Anspruch 21,
dadurch gekennzeichnet,
daß die Enzyme Restriktionsendonukleasen sind.

23. Träger für die Sequenzierung von Nukleinsäuren mit einer Oberfläche, die an einer Vielzahl von vorbestimmten Bereichen immobilisierte Hybridisierungssonden enthält, wobei die Hybridisierungssonden in einzelnen Bereichen jeweils eine unterschiedliche Basenfolge mit einer vorbestimmten Länge aufweisen, wobei die Hybridisierungssonden neben variablen Abschnitten der Länge n einen oder mehrere für zumindest einen Teil der Sonden festgewählte Abschnitte der Länge p aufweisen können.

24. Träger nach Anspruch 23,
dadurch gekennzeichnet,
daß er ein mikrofluidischer Träger ist.

25. Verwendung des Trägers nach Anspruch 23 oder 24 in einem Verfahren zur Sequenzierung von Nukleinsäuren.

26. Verwendung eines Verfahrens nach einem der Ansprüche 1 bis 22 oder des Trägers nach Anspruch 23 oder 24 zur Sequenzierung von Genomen, Chromosomen, Plasmiden, BACs oder/und YACs.

27. Verwendung eines Verfahrens nach einem der Ansprüche 1 bis 22 oder des Trägers nach Anspruch 23 oder 24 zur Transkriptomsequenzierung.

28. Verwendung eines Verfahrens nach einem der Ansprüche 1 bis 22 oder des Trägers nach Anspruch 23 oder 24 zur Identifizierung von Polymorphismen.